Jared Yu

Test 1

1. Assume that a linear regression model has three regressors.
   1. Derive a test statistic to test the null hypothesis that their respective regression coefficients are all zero.

Ans:

A linear regression model with regressors appears as follows,

A hypothesis test that is checking if the regression coefficients are all zero would appear as follows,

The next step is to derive a test statistic for the above hypothesis. The test statistic used is the test statistic which appears as follows,

where

In the above, it is also the case that there is a total of observations, for , . Another assumption is that exists, in other words is invertible. This requires that the columns of are linearly independent. Here, is the design matrix that includes the variables , , and , in addition to a column of 1’s to match with the intercept term .

The textbook also states that

where and is the “centered” model matrix. The can be seen as follows,

It can be seen that each of the columns have their values subtracted by their respective sample means.

If the test statistic is large, then that indicates that is large. This implies that it is likely that at least one of , for . The test statistic actually follows a noncentral distribution with degrees of freedom and that has a noncentrality parameter

The noncentrality parameter indicates that the test statistic should be large if at least one for . To test the hypothesis, then the null can be rejected if , where is the confidence level for the hypothesis test.

* 1. If the null hypothesis in a) is rejected, would this mean that the regression model has value for prediction?

Ans:

The rejection of the null hypothesis stated above implies that at least one of the regressors, , , and contribute significantly to the model. This implies that for example out of , , and at least one of them is a useful regression coefficient at predicting . This does not necessarily mean that the current regression model is good. For example, it is possible that only is nonzero and the others are roughly zero. In that case, it’d possibly make sense to slim down the model to a simple linear regression model involving only as the regressor. In such a case, the regression model with regressors could possibly provide some value, but it would definitely benefit from some refinement.

* 1. Can the importance of each regressor be assessed by simply looking at the respective magnitudes of the -statistics?

Ans:

If the hypothesis test in part b) is truly used and the result is such that the null is rejected, then the next step is to try and figure out which of the coefficients are worth keeping in the model. In such a problem, it’d help to create individual hypothesis tests as follows,

This would test each of the variables , , and to see if they can be removed from the model. If we fail to reject the null hypothesis for any of the , , then that variable can be deleted from the model.

The test statistic is as follows,

where is the diagonal element of that corresponds to (*Note*: it is assumed that is calculatable, since it was used previously to calculate the terms). The null hypothesis is rejected in the case if at confidence level (*Note:* Here, is the Student’s -distribution with degree of freedom . Also, here .). Furthermore, this test is looking at the contribution of the variable , when the other variables are included in the model. Therefore, it is not so straightforward to say that a regressor is good simply because it has a large test statistic . It must also be understood that it is within the context of all the other regressors included.

**State the assumptions for each step of your discussion or derivation in a), b), c).**

1. In a simple linear regression analysis where the regressor is non-random, we have random errors and residuals.
   1. What are the differences between residuals and random errors?

Ans:

The simple linear regression is as follows,

where the intercept and the slope are unknown constants and is a random error component. The errors are assumed to have mean zero and unknown variance . Additionally, an assumption is that the errors are uncorrelated. This term, , helps to complete the linear relationship between the and observations in the data. In general, there will not be a perfect linear fit, and so the random error helps to compensate for the difference between and .

Then, for pairs of data the sample regression model is,

Applying the method of least squares, the intercept, , and slope, , can be estimated with

where and . The fitted simple linear regression model can be written as,

After having calculated the fitted model, the residual is calculated as follows,

It is the difference between the observed values and the fitted value for that sample. These residuals are helpful in the checking the model adequacy. By observing the behavior of the residuals, it can be a way to see if there is a departure from the underlying assumptions for the model.

Compared to the random error, the residual is like the realization and the random error is the random variable. For example, if there is a random variable that represents the result of a die roll, then could be the realization of that random variable after a single roll of the die. So, from the theoretical model, there is , but in the application of the model in a sample, there is . It can be seen then that is like an unknown population variable and is like an observable variable based on a sample of data.

* 1. Derive the variance of residual and the variance of random error.

Ans:

The variance of the residual is as follows:

*Note:* , the reason is that the are fixed and so they are considered constants and so the term can come out of . The denominator can be treated as a constant term. The numerator can be rewritten as follows,

Going back to the previous equation…

*Note: An important fact used above is that the terms are uncorrelated and so the covariance between and for is 0.*

The above equation shows that the variance of the residuals can be written as follows,

In the above derivation, some assumptions used are that the terms are independent which is important within the variance and covariance functions. Also, the terms are treated as regular constants and not as random variables.

In simple linear regression, some assumptions are that and are unknown constants, while the random error is assumed to have mean zero and unknown variance . The random error is also often said to follow a normal distribution, . This is an idealized version of the simple linear regression model and it is not always the case that this property holds within a sample of data. So, it makes sense to say that the variance of the random error is , but it can’t be shown in a mathematical manner.

* 1. Now assume is random. Derive the variance of residual.

Ans:

In the case that is random, there are different steps required in order to find out the variance of the residual. Starting from , it is possible to calculate up to the following step,

The reason it stops here is that since the terms are random, they can’t be easily separated within the term like they were before. So, the derivation for the variance of the residual will begin from another point. It will try the variance based on the formula for the law of total variance.

*Note: Here, is a random variable.* Let us first look at .

Now, let us look at .

*Note:* An assumption is that and are jointly distributed, but the joint distribution is unknown. Also, the are independent random variables whose probability distribution doesn’t include , , or . Also, let represent the variance of the random variable .

Next, let us look at . From part b), it has been shown that this evaluates to,

Then, taking the expectation we have,

The expectation can’t be simplified further, since there is not enough information about , for example what sort of distribution it follows.

Finally, going back to and combining the terms,

This formula is quite similar to the previous version in part b), however it includes some additional complexity due to being a random variable.

**State the assumptions for each step of your discussion or derivation in a), b), c).**

1. In a multiple linear regression analysis, the response variable is studied with two non-random variables and . This regression model that is fitted to the data on of subjects is given by

where , , is a random error with mean zero and variance (its value is unknown), and the weights and have known values.

* 1. Derive the ordinary least-squares estimators for as functions of , not .

Ans: (*Note: Some of the answer is borrowed from a discussion question that utilizes the same methodology to derive the OLS estimators in a multiple linear regression model with regressors.*)

Let be the design matrix for , the data matrix with dimensions . It has dimensions , where , since it includes the column of 1’s in the first position. Let be the vector of the observations. An assumption for OLS is that , , , and that the errors are uncorrelated.

Using the least-squares normal equations, the OLS estimates of for , are as follows,

This formula is based on the least-squares normal equations, , which can be found by minimizing the following,

An assumption for this is that exists, which is possible if the regressors (i.e., and ) are linearly independent. The next step is to find . Let

then

*Note: To simplify notation, allow to be shortened to .*

Furthermore, let . The following steps will attempt to find the inverse of . The row operations will be abbreviated with R1, R2, and R3.

Divide R1 by n:

R2 – R1\*:

R3 – R1\*

Divide R2 by

R3 – R2 \*

Divide R3 by

R2 – R3 \*

R1 – R3\*:

R1 – R2\*

The following has been shown:

Where

Then, from equation :

From this result it follows that for **:**

Before rewriting the above terms in the form of and , the separate variables will be changed first.

Similarly for :

In the above variables,

Given the formulas for , , and , they will now be rewritten in terms of . However, there are some common terms in the formulas that will first be written out. These terms include: , , , and .

First term:

Second term:

Third term:

Last term:

Therefore the terms can be rewritten as follows:

So, it has been shown that:

where

* 1. Derive a statistical test for testing statistical significance of the regression model for , not for .

Ans: (*Note: Answer partially used from Problem 1.*)

A hypothesis test that is checking if there is a linear relationship between the response and any of the regressor variables , appear as follows,

The test statistic for the above hypothesis, , appears as follows,

where

In the above, it is also the case that there is a total of observations, for , . Another assumption is that exists, in other words is invertible. This requires that the columns of are linearly independent. Here, is the design matrix that includes the variables and in addition to a column of 1’s to match with the intercept term .

The problem however is asking that the test statistic be written in terms of rather than . Therefore, some changes need to be made to the variables for and . It has already been shown in part a) how can be written in terms of rather than . Therefore, that step won’t be repeated. Also, the formulas for and includes a few more variables such as: , , , and . The first three terms, , , and , do not depend on , and so they remain unchanged for this part of the problem. However, needs to be expressed in terms of ’s rather than ’s. Denote a new matrix, , as follows,

Now, the test statistic, , can be written as follows,

To test the hypothesis, then the null can be rejected if , where is the confidence level for the hypothesis test and is the -distribution with degrees of freedom and . If we reject the null, then we can say that at least one of the regressors or (i.e., or ) contributes significantly to the model.

**State the assumptions for each step of your discussion or derivation in a), b), c).**

1. Suppose that subjects give data following the true model

where and are non-random regressors, and is the random error.

Now suppose that the reduced model

is also fitted to the same data to obtain the least-squares estimator .

Discuss by mathematical arguments whether this least-squares estimator is biased for . If yes, discuss by mathematical arguments the conditions under which is unbiased for .

Ans:

To show whether or not is unbiased, it must be shown that . Using the simple linear regression model, denote the OLS estimators of the reduced model with and . Then, these estimators can be shown to be

where . This is done by optimizing the following formula,

*Note:* Let represent the sample from the variable, where . It has been shown previously in problem 2 b) how can be derived.

Now, let us analyze the expected value of :

*Note: The assumption above is that is non-random and so it can be treated as a constant term for the expectation.*

*Note: An assumption used above is that* .

*Note: An assumption used above is that , , and that is nonzero. The first two will be shown briefly.*

So, from , it can clearly be seen that , so is biased. There are special cases where however. It can be seen that if either or evaluate to 0, then . An example would be if and were identical, or some other way where the numerator .

**State the assumptions for each step of your discussion or derivation.**